Additional paper for “Machine Learning for Protein Classification”

Paper reference
Marco Cuturi, Jean-Philippe Vert, “The Context-Tree Kernel for Strings”

Abstract
Cuturi and Vert present a novel class of string kernels that are derived using context-tree weighing algorithm, a popular string compression algorithm. The intuition behind the method is that if applying a string compression algorithm on two input strings—X and Y—behaves similarly when applying the same string compression algorithm on the concatenation of the two strings—XY—it can be said that the strings share similar properties. On the other hand, if the concatenation cannot be easily compressed, one can say that the strings are dissimilar. This can be thought of in terms of coding redundancy between X and Y with respect to XY.

Discussion
The kernels produced by the context-tree algorithm can be categorized under the class of mutual information kernels, which is a good representation for a meaningful measure of similarity between two sequences. And since the context-tree weighing algorithm is a general algorithm rooted in information theory, the method is general enough to be efficient on different datasets with excessive tuning.

Broadly speaking, existing kernels can be classified into two main categories: 1) those which are fast to compute, such as the spectrum and mismatch kernel; and 2) those which improve the accuracy of the classifier, such as the Fisher, pairwise, and local alignment kernel. The drawback with the latter category of kernels is that the kernel-finding method requires prior biological knowledge and that the kernels are slow to compute. Because the context-tree algorithm that the authors present in the paper makes no prior assumptions about the nature of the input sequence, it has the advantage of being both fast and generic. Thus, the authors position their method as the one to use in scenarios where one needs to compare, fairly quickly, sequences for which little prior knowledge is available.